Short Communication

Impact of Misclassification in Genotype-Exposure Interaction Studies: Example of *N*-Acetyltransferase 2 (*NAT2*), Smoking, and Bladder Cancer

Anne C. Deitz,[1] Nathanial Rothman,[2] Timothy R. Rebbeck,[1] Richard B. Hayes,[2] Wong-Ho Chow,[2] Wei Zheng,[3] David W. Hein,[4] and Montserrat García-Closas[2]

[1]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania; [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland; [3]Health Services Research, Vanderbilt University, Nashville, Tennessee; and [4]Department of Pharmacology and Toxicology and James Graham Brown Cancer Center, University of Louisville School of Medicine, Louisville, Kentucky

Abstract

Errors in genotype determination can lead to bias in the estimation of genotype effects and gene-environment interactions and increases in the sample size required for molecular epidemiologic studies. We evaluated the effect of genotype misclassification on odds ratio estimates and sample size requirements for a study of NAT2 acetylation status, smoking, and bladder cancer risk. Errors in the assignment of NAT2 acetylation status by a commonly used 3-single nucleotide polymorphism (SNP) genotyping assay, compared with an 11-SNP assay, were relatively small (sensitivity of 94% and specificity of 100%) and resulted in only slight biases of the interaction parameters. However, use of the 11-SNP assay resulted in a substantial decrease in sample size needs to detect a previously reported NAT2-smoking interaction for bladder cancer: 1,121 cases instead of 1,444 cases, assuming a 1:1 case-control ratio. This example illustrates how reducing genotype misclassification can result in substantial decreases in sample size requirements and possibly substantial decreases in the cost of studies to evaluate interactions. (Cancer Epidemiol Biomarkers Prev 2004;13(9):1543–6)

Introduction

Germline genotype information is often used as a surrogate measure of metabolic phenotype in molecular epidemiologic studies. Metabolic phenotyping assays are generally more time-consuming, more expensive, and not suitable for studies employing samples collected after disease diagnosis and treatment. For enzymes such as *N*-acetyltransferase 2 (NAT2), genotype can predict phenotype with a high degree of accuracy (1, 2). However, this requires that all relevant SNPs and/or alleles for the population under study be analyzed (3).

At the time of article submission, there were 29 reported *NAT2* alleles (http://www.louisville.edu/medschool/pharmacology/NAT.html) encoding proteins with varying degrees of acetylation capacity. Each of the 29 *NAT2* alleles possesses a combination of one to four SNPs at 13 sites within the 870-bp coding region. The majority of studies investigating the relationship between *NAT2* genotype and disease risk use PCR-based assays that detect only three SNPs (C481T, G590A, and G857A) to infer NAT2 acetylation status. When none of these SNPs are present, wild-type *NAT2*4*, a high-activity (rapid) allele, is designated (4). Although several *NAT2* SNPs are in linkage disequilibrium, assessment of only these three SNPs results in the misclassification of the following *NAT2* low-activity (slow) alleles (*NAT2*5C*, *NAT2*5D*, *NAT2*14A*, *NAT2*14B*, *NAT2*14E*, *NAT2*14F*, *NAT2*14G*, *NAT2*17*, and *NAT2*19*) as *NAT2*4*, a high-activity (rapid) allele. Additionally, *NAT2*11* and *NAT2*12C*, high-activity alleles, would be misassigned as *NAT2*5B*, a low-activity allele formerly designated as *M1* (see Table 1 for allele descriptions).

Nondifferential misclassification of binary genetic or exposure factors biases odds ratio (OR) estimates toward the null hypothesis and results in decreased statistical power (5). In this article, we illustrate the impact of genotype misclassification on OR estimates and sample size requirements for detecting genotype-exposure interaction. For this purpose, we used an example of NAT2 acetylation status and smoking interaction on bladder cancer risk.

Methods

NAT2 rapid (''R''), intermediate (''I''), and slow (''S'') acetylator phenotypes were determined for an

Received 4/14/03; revised 4/11/04; accepted 4/20/04.

Grant support: National Cancer Institute grant CA34627.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Three new SNPs have been identified in the human *NAT2* coding-region, resulting in 7 additional *NAT2* alleles. Of these 16 SNPs, we continue to recommend screening for the seven most common: G191A, C282T, T341C, C481T, G590A, A803G, and G857A.

Requests for reprints: Montserrat García-Closas, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892. Phone: 301-435-3981; Fax: 301-402-0916. E-mail: garciacm@exchange.nih.gov

institutional review board–approved case-control study of stomach cancer (6) using a previously described PCR-RFLP assay (7). This assay, developed by Doll et al., can detect 11 SNPs that determine all 26 allele variants reported when this project began. The assay requires initial amplification of the entire *NAT2* coding region followed by three sets of double restriction enzyme digests: *Msp*I/*Kpn*I to detect G191A, A434C, and C481T; *Taq*I/*Bam*HI to detect T111C, G590A, C759T, and G857A; and *Fok*I/*Dra*III to detect C282T and A845C. T341C and A803G are detected with nested PCR reactions and subsequent enzyme digests.

NAT2 phenotypes were also assigned by assuming that a 3-SNP (C481T, G590A, and G857A) rather than the 11-SNP assay had been used. In both instances, individuals were classified as ''R'' if they possessed two high-activity alleles (*NAT2*4*, *NAT2*11*, *NAT2*12A*, *NAT2*12B*, *NAT2*12C*, *NAT2*13*, and *NAT2*18*), ''I'' if they possessed one of these alleles, and ''S'' if they possessed none. All genotype assignments were blinded to case-control status.

To compare ''R,'' ''I,'' and ''S'' phenotype assignments made by the 3-SNP assay relative to the 11-SNP assay (gold standard), a 3 × 3 misclassification table was created for controls from the case-control study of stomach cancer. Although recent data suggest that ''R'' and ''I'' are likely separate phenotypes (8-10), for simplicity, NAT2 acetylation status was dichotomized into ''S'' and ''I/R'' groups, and *NAT2* misclassification probabilities (e.g., sensitivity and specificity) were determined. Given this bimodal phenotype model, misclassification of ''I'' as ''R'' or ''R'' as ''I'' could not be evaluated. To confirm that sensitivity and specificity values were not unique to this population, sensitivity and specificity were determined as described above for controls from a case-control study of breast cancer comprised of Caucasian women from Iowa (7) and an unpublished case-control study of prostate cancer comprised of 45% African-Americans.

Estimates for prevalence of smoking, prevalence of NAT2 acetylation status, OR of smoking (OR_E), OR of NAT2 acetylation status (OR_G), and the multiplicative genotype-smoking interaction parameter (ψ) were based on data from previously published European studies of NAT2, smoking, and bladder cancer that used the 3-SNP assay (11). Sensitivity and specificity were used to calculate expected parameters in the absence of misclassification (12). The expected values for these five parameters using the 11-SNP assay (gold standard) were calculated using formulas described in Garcia-Closas et al. (5). Sample sizes for these genotype-exposure interaction studies were estimated using the POWER software available at http://dceg.cancer.gov/POWER/.

Results

In all three case-control studies, the most commonly occurring alleles among controls were *NAT2*5B*>*NAT2*6A*>*NAT2*4* (Table 1). Not surprisingly, allele distribution was most similar for the two Caucasian populations, although the *NAT2*5A* allele frequency was higher among American Caucasians than European Caucasians. *NAT2*12*, *NAT2*13*, and *NAT2*14* allele cluster frequencies were much higher among prostate controls (45% African Americans) than in the other two populations studied.

As shown in Table 2, agreement between the two genotyping assays for assigning ''R,'' ''I,'' and ''S'' phenotypes was very high among controls in the case-control study of stomach cancer. Relative to the 11-SNP assay, the proportion of individuals correctly classified as a slow acetylator by the 3-SNP method (i.e., sensitivity) was 94% (95% CI, 89–96%), whereas the proportion of individuals correctly classified as a rapid or intermediate acetylator by the 3-SNP method (i.e., specificity) was 100% (95% CI, 98–100%). Sensitivity and specificity values were comparable among controls from the breast cancer study (96% and 100%, respectively). Sensitivity was much lower (83%) for the multiracial prostate cancer controls but increased to 93% when the G191A SNP was added to the assay (data not shown). This SNP is unique to the *NAT2*14* cluster, common among African-American and Hispanic populations (4). Interestingly, of the 16 acetylator phenotypes that were misclassified in the stomach cancer controls, all were due to the *NAT2*5C* (T341C, A803G) allele, whereas 94% of the misclassification in the breast cancer controls was due to *NAT2*5C* (data not shown). In both of these Caucasian case-control studies, the *NAT2*5C* allele frequency was ~2% among controls.

Based on the estimates determined from a recent meta-analysis (11) of NAT2, smoking, and bladder cancer (60% prevalence of smoking and 60% prevalence of slow acetylators, $OR_E = 3.0$, $OR_G = 1.5$, $\psi = 1.65$), 1,444 cases and 1,444 controls would be required detect a genotype-smoking interaction OR of 1.65 at 80% power and $\alpha = 0.05$. After adjusting these parameters for sensitivity and specificity, the joint effects OR remained practically unchanged (observed 3.57 versus expected in the abscence of misclassification 3.63), but ψ increased from 1.65 to 1.78. Thus, in the absence of genotype misclassification (i.e., using the 11-SNP assay rather than the 3-SNP assay), sample size to detect genotype-smoking interaction would have been reduced to 1,121 cases and 1,121 controls. This corresponds to a 22% decrease in sample size.

Discussion

Multiple sources of bias may exist in epidemiologic studies investigating genotype-exposure interaction. Although most investigators recognize the need for improving the accuracy of exposure assessment, less attention has been given to reducing genotype misclassification because genotypes are usually measured with a higher level of accuracy than environmental exposures. One obvious way to reduce genotype misclassification is by employing validated laboratory assays. This eliminates errors associated with poor assay design such as amplification of a pseudogene and incomplete restriction enzyme digests.

Another way that misclassification can be reduced is by determining all SNPs that are relevant to inferred phenotype, as we have shown in this example. Similarly, it is important to screen for all SNPs that are relevant to the race/ethnicity of the sample population. The 3-SNP *NAT2* assay was designed to detect the most frequently

Table 1. *NAT2* allele distribution among controls from case-control studies of breast, prostate, and stomach cancers

Allele	Nucleotide substitution(s)	Breast, n = 387 [n Alleles (%)]	Stomach, n = 414 [n Alleles (%)]	Prostate, n = 149 [n Alleles (%)]
NAT2*4	None	187 (24.2)	219 (26.4)	64 (21.5)
*NAT2*5A*	T341C, C481T	20 (2.6)	7 (0.85)	5 (1.7)
*NAT2*5B*	T341C, C481T, A803G	318 (41.1)	309 (37.3)	104 (34.9)
*NAT2*5C*	T341C, A803G	17 (2.2)	17 (2.1)	9 (3.0)
*NAT2*5D*	T341C	—	—	—
*NAT2*5E*	T341C, G590A	—	—	1 (0.34)
*NAT2*5F*	T341C, C481T, C759T, A803G	—	—	—
*NAT2*6A*	C282T, G590A	206 (26.6)	251 (30.3)	66 (22.1)
*NAT2*6B*	G590A	—	1 (0.12)	1 (0.34)
*NAT2*6C*	C282T, G590A, A803G	—	—	—
*NAT2*6D*	T111C, C282T, G590A	—	—	—
*NAT2*7A*	G857A	—	—	—
*NAT2*7B*	C282T, G857A	15 (1.9)	19 (2.3)	8 (2.7)
*NAT2*10*	G499A	ND	ND	ND
NAT2*11	C481T	—	—	—
NAT2*12A	A803G	3 (0.4)	4 (0.48)	6 (2.0)
NAT2*12B	C282T, A803G	—	—	4 (1.3)
NAT2*12C	C481T, A803G	—	—	2 (0.67)
NAT2*13	C282T	7 (0.9)	1 (0.12)	14 (4.7)
*NAT2*14A*	G191A	—	—	3 (1.0)
*NAT2*14B*	G191A, C282T	1 (0.1)	—	11 (3.7)
*NAT2*14C*	G191A, T341C, C481T, A803G	—	—	—
*NAT2*14D*	G191A, C282T, G590A	—	—	—
*NAT2*14E*	G191A, A803G	—	—	—
*NAT2*14F*	G191A, T341C, A803G	—	—	—
*NAT2*14G*	G191A, C282T, A803G	—	—	—
*NAT2*17*	A434C	—	—	—
NAT2*18	A845C	—	—	—
*NAT2*19*	C190T	ND	ND	ND

NOTE: Alleles were assigned using a PCR-based assay that detects 11 SNPs and can therefore distinguish among 26 *NAT2* allele variants. Alleles in boldface are high-activity (rapid) alleles, whereas all others are low-activity (slow) alleles. ''Intermediate'' acetylator phenotype is assigned when an individual possesses one ''slow'' and one ''rapid'' allele. It should be noted that *NAT2*10* phenotype is unknown. *ND,* Our assay does not detect the G499A or C190T SNPs and thus cannot distinguish these alleles.

occurring *NAT2* alleles in Caucasian populations, so it was no surprise that its sensitivity was high among our Caucasian controls. The 3-SNP assay, however, performs more poorly in other racial/ethnic groups as shown in the control population that included a high percentage of African-Americans. Based on 11-SNP screening of 950 alleles, we found that seven SNPs (G191A, C282T, T341C, C481T, G590A, A803G, and G857A) explained 100% of the alleles that were detected. Therefore, we recommend that these seven SNPs be screened in Caucasian and African-American populations to accurately infer NAT2 acetylator phenotype. A TaqMan assay, which costs less than one dollar per SNP, has recently been developed for this purpose (13). It is important to note that the number of SNPs that need to be determined to attain high accuracy in phenotype assignments may vary depending on the ethnic background of the population under study because of SNP prevalence across ethnic groups. See http://snp500cancer.nci.nih.gov for useful information on *NAT2* SNP frequencies in four subpopulations; unfortunately, comprehensive *NAT2* SNP screening has not been done in many ethnic groups. Until then, we recommend that at least seven *NAT2* SNPs be screened in most populations, especially given the relatively low cost of genotyping and the potential for population admixture.

Table 2. Concordance between two *NAT2* genotyping assays among controls from case-control studies of stomach, breast, and prostate cancers

3-SNP assay	11-SNP assay			
	S	I	R	Total
Stomach				
S	209	0	0	209
I	13	156	0	169
R	1	2	33	36
Total	223	158	33	414
Breast				
S	204	0	0	204
I	9	142	0	151
R	0	9	23	32
Total	213	151	23	387
Prostate				
S	62	1	0	63
I	12	50	1	63
R	1	7	15	23
Total	75	58	16	149

Although our 11-SNP assay is comprehensive, allele (or haplotype) assignment can sometimes be ambiguous. For example, an individual who is typed as a heterozygote at nucleotides 341 and 803 may be a *NAT2*5D/NAT2*12A* if both SNPs reside on separate alleles or a *NAT2*5C/NAT2*4* if both SNPs are located on the same allele. Because *NAT2* polymorphisms are well characterized, it is possible to collapse resulting genotypes into inferred phenotype categories. In this case, both genotypes result in the assignment of ''I'' phenotype. When function is largely unknown, however, correct allele/haplotype assignment is critical. Recent advances in high-throughput genotyping should facilitate comprehensive SNP screening of other highly polymorphic loci, such as *NAT1* and *CYP2D6*.

Our results indicate that, despite relatively small errors in NAT2 phenotype assignments and small biases in OR estimates, substantial decreases in sample size required to detect genotype-exposure interaction can be attained using the 11-SNP *NAT2* genotyping assay rather than the 3-SNP assay. Given the expense associated with enrolling subjects in molecular epidemiologic studies, reducing genotype misclassification is likely to result in substantial reduction in study costs. In addition, reducing genotype misclassification will reduce the bias in the estimated parameters.

References

1. Cascorbi I, Brockmoller J, Mrozikiewicz PM, Muller A, Roots I. Arylamine *N*-acetyltransferase activity in man. Drug Metab Rev 1999;31:489–502.
2. Gross M, Kruisselbrink T, Anderson K, et al. Distribution and concordance of *N*-acetyltransferase genotype and phenotype in an American population. Cancer Epidemiol Biomarkers Prev 1999;8:683–92.
3. Rothman N, Stewart WF, Caporaso NE, Hayes RB. Misclassification of genetic susceptibility biomarkers: implications for case-control studies and cross-population comparisons. Cancer Epidemiol Biomarkers Prev 1993;2:299–303.
4. Bell DA, Taylor JA, Butler MA, et al. Genotype/phenotype discordance for human arylamine *N*-acetyltransferase (NAT2) reveals a new slow-acetylator allele common in African-Americans. Carcinogenesis 1993;14:1689–92.
5. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. Cancer Epidemiol Biomarkers Prev 1999;8:1043–50.
6. Lan Q, Rothman N, Chow WH, et al. No apparent association between NAT1 and NAT2 genotypes and risk of stomach cancer. Cancer Epidemiol Biomarkers Prev 2003;12:384–6.
7. Deitz AC, Zheng W, Leff MA, et al. *N*-acetyltransferase-2 genetic polymorphism, well-done meat intake, and breast cancer risk among postmenopausal women. Cancer Epidemiol Biomarkers Prev 2000;9:905–10.
8. Hein DW, Doll MA, Fretland AJ, et al. Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. Cancer Epidemiol Biomarkers Prev 2000;9:29–42.
9. Fretland AJ, Leff MA, Doll MA, Hein DW. Functional characterization of human *N*-acetyltransferase 2 (*NAT2*) single nucleotide polymorphisms. Pharmacogenetics 2001;11:207–15.
10. Le Marchand L, Hankin JH, Wilkens LR, et al. Combined effects of well-done red meat, smoking, and rapid *N*-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. Cancer Epidemiol Biomarkers Prev 2001;10:1259–66.
11. Marcus PM, Hayes RB, Vineis P, et al. Cigarette smoking, *N*-acetyltransferase 2 acetylation status, and bladder cancer risk: a case-series meta-analysis of a gene-environment interaction. Cancer Epidemiol Biomarkers Prev 2000;9:461–7.
12. Flegal KM, Brownie C, Haas JD. The effects of exposure misclassification on estimates of relative risk. Am J Epidemiol 1986;123:736–50.
13. Doll MA, Hein DW. Comprehensive human NAT2 genotype method using single nucleotide polymorphism-specific polymerase chain reaction primers and fluorogenic probes. Anal Biochem 2001;288:106–8.